# IMPROVED BANDWIDTH UTILIZATION IN A PPRC SYSTEM

## TECHNICAL FIELD

[1] The present invention relates generally to peer-to-peer remote copy (PPRC) storage systems and, in particular, to increasing the efficiency with which write operations are performed between a primary storage control unit and a secondary storage control unit.

## BACKGROUND ART

[2] Data integrity is a critical factor in large computer data systems. Consequently, backup data storage systems have been developed to prevent the loss of data in the event of various types of failures. One such backup system, known as "peer-to-peer remote copy" (PPRC), maintains a separate, consistent copy of data. As illustrated in Fig. 1, in a PPRC system 100, data generated by a host device 110 is stored on a primary storage unit 120. A copy of the data is also transmitted synchronously, such as over a fibre channel network 130, and stored on a secondary storage unit 140. Because of the flexibility of network interconnections, the primary and secondary units 120 and 140 may be physically located remote from the host 110. And, for additional data security, the primary and secondary units 120 and 140 may be (but need not be) physically located distant from each other, thereby reducing the likelihood of a single disaster simultaneously harming both the primary and secondary units 120 and 140. It will be appreciated that the primary and secondary units 120 and 140 may be the same physical unit, divided logically into two.

[3] Typically, a block of data to be copied to the secondary unit 140 is first received by the primary unit 120 as a plurality of substantially equal-size units or tracks. A first track is then associated with a tack, thread or other operating system unit of execution, such as a task control block (TCB) and transferred from the primary unit to the secondary over one of the links. Upon successful receipt of the track, the secondary replies with a "complete" status message. The TCB is then released and the next track may then be associated with the TCB and transferred to the secondary. The process is repeated until all of the tracks have been successfully



Docket: TUC920030143US1
Express Mail Label: EV303489229US

transferred to the secondary. Such a serial operation may fail to take full advantage of the bandwidth available between the primary and secondary storage units.

[4] It is also possible to transfer tracks in a "piped" fashion in which the transfer of a second track begins before the status message has been received from the secondary unit confirming successful receipt of the first track. However, there may be insufficient resources in the primary unit to complete the transfer of all of the tracks of a block of data in such a fashion. Additionally, if too many resources are allocated to piping tracks, other operations may not have sufficient resources and may be delayed.

[5] Consequently, it remains desirable to provide a more efficient and flexible process for transferring blocks of data from a primary storage unit to a secondary storage unit.

## SUMMARY OF THE INVENTION

[6] The present invention provides method, system and computer program product to improve the efficiency of data transfers in a PPRC environment. A block of data to be transferred is divided into tracks. Each track is allocated to a data mover task control block (TCB) with a master TCB being assigned to supervise the data mover TCBs. The tracks are then transferred from the primary storage controller to the secondary controller in a piped fashion over a link coupling the primary and secondary storage controllers. However, the usage of resources is monitored by a resource management algorithm and, if too many TCBs are being used for a transfer (or if the supply of data mover TCBs is exhausted), the transfer is automatically switched whereby the master TCB now serves as the data mover TCB for the remaining tracks.

[7] In addition, the various links coupling the primary and secondary storage controllers are monitored to determine which link will provide the fastest transfer. If, during a transfer of tracks over one link, a faster link is identified, the transfer may be switched to the second, faster link.

[8] Thus, the efficiency of a transfer of data is improved and the utilization of the bandwidth of links is similarly improved.

## BRIEF DESCRIPTION OF THE DRAWINGS

[9] Fig. 1 is a block diagram of an exemplary PPRC data storage system;

[10] Fig. 2 is a block diagram of a data storage system in which the present invention may be implemented;

[11] Fig. 3A is a flow chart of a method of the present invention;

[12] Fig. 3B is a flow chart of a method of the present invention; and

[13] Fig. 4 illustrates tracks into which a block of data has been divided, a master TCB and data mover TCBs with which the tracks are transferred.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[14] Fig. 2 is a block diagram of a data storage system 200 in which the present invention may be implemented. The system 200 may be a PPRC system in which a host device 210 transmits data to a primary storage controller 220 for storage on a storage device (such as an array of disk dives) 228. Additionally, the primary storage controller 220 transfers a copy of the data to a secondary storage controller 240, such as over a network 230, for storage on a storage device 248. The secondary storage controller 240 and storage device 248 are preferably, but not necessarily, located remote from the primary storage controller 220. The primary storage controller 220 further includes a processor 222 for executing instructions relating to the transfer of the copies of data to the secondary storage controller 240. The primary storage controller 220 also includes a memory 224, portions of which are allocated to tasks, threads or other operating system units of execution, such as task control blocks (TCBs). The task control blocks may include master TCBs 410 And a plurality of data mover TCBs 420 associated with each master TCB 410.

[15] The secondary storage controller also includes a processor 242 to execute instructions relating to the receipt and ultimate storage of data.

[16] The network 230 preferably includes a plurality of links 232, 234 and 236 through which data and messages may be transferred between the primary and secondary storage controllers 220 and 240. Although three such links 232, 234 and 236 are shown in Fig. 2, it will be appreciated that any number of links may be provided. The

network 230 and the links 232, 234 and 236 are illustrated as directly connecting the primary storage controller 220 and the secondary storage controller 230. However, it will be appreciated that such a depiction is merely for clarity in the Fig. and that the illustrated network 230 and links 232, 234 and 236 represent any path, direct or indirect, by which the primary storage controller 220 and the secondary storage controller 230 may be coupled.

[17] In operation, a block of data 400 (Fig. 4) at the host 210 is divided into substantially equal-size tracks which are transferred to the primary storage controller 220 to be stored in the storage device 228. Referring to the flow chart of Fig. 3A, as tracks arrive at the primary storage controller 220, memory space 224 in the primary storage controller 220 is allocated to a master TCB 410 (step 300) and to a plurality of data mover TCBs 420 (step 302). As part of a PPRC write command issued by the primary storage controller 220 to commence a write operation to transfer the block of data 400 to the secondary storage controller 240 (step 304), the data mover TCBs 420 are placed in a queue associated with the master TCB 410 (step 306). As tracks arrive from the host 210, each is associated with one of the data mover TCBs 420 (step 308).

[18] The master mover TCB 410 directs that a first track 402 be transferred using a first data mover TCB 422 over a first link 232 through which the primary and secondary storage controllers 220 and 240 are coupled (step 310). The first data mover TCB 422 is released (step 312) and may be used during the transfer of another block of data. Prior to receiving an acknowledgment from the secondary storage controller 240 that the first track 402 was successfully received, the next track 404 is transferred using a next data mover TCB 424 (step 314) over the same link 232; the second data mover TCB is then released (step 316).

[19] Usage of the data mover TCBs 420 is monitored by a resource management algorithm (step 318). If a sufficient number of data mover TCBs 420 are available to complete the transfer of the block of data 400, transfer of the remaining tracks continues as before until all of the tracks have been transferred (step 320). However, if an insufficient number of data mover TCBs 420 are available or if the supply of data mover TCBs 420 is exhausted, the master TCB 410 becomes a data

mover TCB (step 322) and the remaining tracks are transferred serially (step 324) with the transfer of a next track delayed until the track has been received (step 326) and an acknowledgment from the secondary storage controller 240 is received (step 328).

[20] At the secondary storage control unit 240, the tracks are received (step 326), an acknowledgment transmitted (step 328) and the tracks are reassembled into the block of data 400 (step 330) which is ultimately stored on the storage device 248 (step 330), thereby completing the write operation. Upon receipt by the primary storage controller 220 from the secondary storage controller 240 that all of the tracks have been successfully received (step 334), the master TCB 410 is released (step 336) and may be used for the transfer of another block of data.

[21] Referring to Fig. 3B, in addition to monitoring usage of the data mover TCBs 420, usage of the links 232, 234 and 236 coupling the primary and secondary storage controllers 220 and 240 is also monitored (step 338) to determine which link may provide the most efficient (such as the fastest) transfer of data. If a link is identified which is more efficient than the link currently being used (step 340), transfer of tracks may be switched to the more efficient link (step 342). Preferably, the switch to a more efficient link occurs after the primary storage controller 220 has received an acknowledgment from the secondary storage controller 240 that a first group or group of tracks has been successfully received (340). Otherwise, the tracks of data may not all arrive in the proper order (some later tracks transferred over the more efficient link, may arrive before some earlier tracks transferred over the original link) and the secondary storage controller 240 will have to rearrange the tracks into the original order. If no more efficient link is identified, transfer of the tracks continues over the original link (step 344).

[22] The objects of the invention have been fully realized through the embodiments disclosed herein. Those skilled in the art will appreciate that the various aspects of the invention may be achieved through different embodiments without departing from the essential function of the invention. The particular embodiments are

illustrative and not meant to limit the scope of the invention as set forth in the following claims.